ARRANGEMENT AND METHOD FOR RECOGNIZING A PREDETERMINED VOCABULARY IN SPOKEN LANGUAGE WITH A COMPUTER

The invention is directed to an arrangement and a method for the

5   recognition of a predetermined vocabulary in spoken language by a computer.

A method and an arrangement for the recognition of spoken language are

known from [1]. Particularly until a recognized word sequence is obtained from a

digitalized voice signal, a signal analysis and a global search that accesses an acoustic

model and a linguistic model of the language to be recognized are implemented in the

10   recognition of spoken language. The acoustic model is based on a phoneme inventory

realized with the assistance of hidden Markov models (HMMs). With the assistance

of the acoustic model, a suitable probable word sequences is determined during the

global search for feature vectors that proceeded from the signal analysis and this is

output as recognized word sequence. The words to be recognized are stored in a

15   pronunciation lexicon together with a phonetic transcription. The relationship is

explained in depth in [1].

For explaining the subsequent comments, the terms that are employed

shall be briefly discussed here.

As phase of the computer-based speech recognition, the signal analysis

20   particularly comprises a Fourier transformation of the digitalized voice signal and a

feature extraction following thereupon. It proceeds from [1] that the signal analysis

ensues every ten milliseconds. From overlapping time segments with a respective

duration of, for example, 25 milliseconds, approximately 30 features are determined

on the basis of the signal analysis and combined to form as feature vector. The

25   components of the feature vector describe the spectral energy distribution of the

appertaining signal excerpt. In order to arrive at this energy distribution, a Fourier

transformation is implemented on every signal excerpt (25 ms time excerpt). The

components of the feature vector result from the presentation of the signal in the

frequency domain. After the signal analysis, thus, the digitalized voice signal is

30   present in the form of feature vectors.

These feature vectors are supplied to the ~~global search~~, a further phase of the speech recognition. As already mentioned, the global search makes use of the ~~acoustic model~~ and, potentially, of the linguistic model in order to image the sequence of feature vectors onto individual parts of the language (vocabulary) present as model.

5   A language is composed of a given plurality of sounds, ~~what are~~ referred to as phonemes, whose totality is referred to as phoneme inventory. The vocabulary is modelled by phoneme sequences and stored in a pronunciation lexicon. Each phoneme is modelled by at least one HMM. A plurality of HMMs yield a stochastic automaton that comprises statusses and status transitions. The time execution of the

10   occurrence of specific feature vectors (even within a phoneme) can be modelled with HMMs. A corresponding phoneme model thereby comprises a given plurality of statusses that are arranged in linear succession. A status of an HMM represents a part of a phoneme (for example an excerpt of 10 ms length). Each status is linked to an emission probability, which, in particular, is distributed according to Gauss, for the

15   feature vectors and to transition probabilities for the possible transitions. A probability with which a feature vector is observed in an appertaining status is allocated to the feature vector with the emission distribution. The possible transitions are a direct transition from one status into a next status, a repetition of the status and a skipping of the status.

20   A joining of the HMM statusses to the appertaining transitions over the time is referred to as trellis. The principle of dynamic programming is employed in order to determined the acoustic probability of a word: the path through the trellis is sought that exhibits the fewest errors or, respectively, that is defined by the highest probability for a word to be recognized.

25   The result of the global search is the output or, respectively, offering of a recognized word sequence that derives taking the acoustic model (phoneme inventory) for each individual word and the language model for the sequence of words into consideration.

[2] discloses a method for speaker adaptation based on a MAP estimate

30   (MAP = maximum a posteriori) of HMM parameters.

According to [2], thus, it is recognized that a speaker-dependent system for speech

recognition normally supplies better results than a speaker-independent system,

insofar as adequate training data are available that enable a modelling of the speaker-

dependent system. However, the speaker-independent system achieves the better

5   results as soon as the set of speaker-specific training data is limited. One possibility

for performance enhancement of both systems, i.e. of both the speaker-dependent as

well as the speaker-independent system for speech recognition, is comprised in

employing previously stored datasets of a plurality of speakers such that a small set of

training data also suffices for modelling a new speaker with adequate quality. Such a

10   training method is called speaker adaptation. In [2], the speaker adaptation is

particularly implemented by a MAP estimate of the hidden Markov model parameters.

Results of a method for recognizing spoken language generally deteriorate

as soon as characteristic features of the spoken language deviate from characteristic

features of the training data. Examples of characteristic features are speaker qualities

15   or acoustic features that influence the articulation of the phonemes in the form of

slurring.

The approach ~~embarked upon in [2]~~ employs for speaker adaptation ~~is comprised in~~ a

"post-estimating" parameter values of the hidden Markov models, whereby this

processing in implemented "offline", i.e. not at the run time of the method for speech

20   recognition.

J. Takami et al., "Successive State Splitting Algorithm for Efficient

Allophone Modeling", ICASSP 1992, March 1992, pages 573 through 576, San

Francisco, USA, discloses a method for recognizing a predetermined vocabulary in

spoken language wherein states are split in a hidden Markov model. The probability

25   density function of the respective states is also split therefor. And object

~~The object~~ of the invention is ~~comprised is specifying~~ to provide an arrangement and

a method for recognizing a predetermined vocabulary in spoken language, whereby,

within

in particular, an adaptation of the acoustic model is accomplished ~~at~~ the run time (i.e.,

"online").

30   ~~This object is achieved according to the features of the independent patent~~

~~claims.~~

in the inventive

For achieving the object, a method for recognizing a predetermined vocabulary in spoken language with a computer is recited wherein a voice signal is determined from the spoken language. The voice signal is subjected to a signal

5 analysis from which feature vectors for describing the digitalized voice signal proceed. A global search is implemented for imaging the feature vectors onto a

language present in modelled form, whereby each phoneme of the language is described by a modified hidden Markov model and each status of the modified hidden Markov model is described by a probability density function. An adaptation of the probability density function ensues such that it is split into a first probability density

5 function and into a second probability density function. Finally, the global search offers a recognized word sequence.

It should be noted that the probability density function that is split

a into a first and into a second probability density function can represent an emission distribution for a predetermined status of the modified hidden Markov model,

10 whereby this emission distribution can also contain a superimposition of a plurality of probability density functions, for example Gauss curves (Gaussian probability density distributions.

A recognized word sequence can thereby also comprise individual sounds or, respectively, only a single word.

15 If, in the framework of the global search, a recognition is affected with a high value for the distance between spoken language and appertaining word sequence determined by the global search, then the allocation of a zero word can ensue, said zero word indicating that the spoken language is not being recognized with adequate quality.

20 By splitting the probability density function, one advantage of the invention is to create new regions in a feature space erected by the feature vectors, these new regions comprising significant information with reference to the digitalized voice data to be recognized and, thus, assuring an improved recognition.

In an embodiment of the invention the probability density

25 function is split into the first and into the second probability density function when the drop off of an entropy value lies below a predetermined threshold.

The splitting of the probability density function dependent on an entropy value proves extremely advantageous in practice.

The entropy is generally a measure of an uncertainty in a prediction of a

30 statistical event. In particular, the entropy can be mathematically defined for

Gaussian distributions, whereby there is a direct logarithmic dependency between the scatter σ and the entropy. *In another embodiment*

~~Another development~~ of the invention ~~is comprised therein that~~ the probability density functions, particularly the first and the second probability density

5   function respectively comprise at least one Gaussian distribution.

The probability density function of the status is approximated by a sum of a plurality of Gaussian distributions. The individual Gaussian distributions are called modes. In the recited method, in particular, the modes are considered isolated from one another. One mode is divided into two modes in every individual split event.

10   When the probability density function was formed of m modes, then it is formed of M+1 modes after the split event. When, for example, a mode is assumed to be a Gaussian distribution, then an entropy can be calculated, as shown in the exemplary embodiment.

An online adaptation is advantageous because the method continues to

15   recognize speech without having to be set to the modification of the vocabulary in a separate training phase. A self-adaptation ensues that, in particular, becomes necessary due to a modified co-articulation of the speakers due to an addition of a new word.

The online adaptation, accordingly, requires no separate calculation of the

20   probability density functions that would in turn be responsible for a non-availability of the system for speech recognition. *In a further embodiment*

~~One development~~ of the invention ~~is comprised therein that~~ identical standard deviations are defined for the first probability density function and for the second probability density function. A first average of the first probability density

25   function and a second average of the second probability density function are defined such that the first average differs from the second average.

This is an example for the weighting of the first and second probability density function split from the probability density function. Arbitrarily other weightings are also conceivable that are to be adapted to the respective application. *In a further embodiment*

30   ~~Further, one development is that~~ the method is multiply implemented in succession and, thus, a repeated splitting of the probability density function ensues.

Developments of the invention derive from the dependent claims. *The above mentioned also achieve accordance with the invention in* ~~Another solution of the~~ object is ~~comprised~~ in ~~specifying~~ an arrangement with a processor unit that is configured such that the following steps can be implemented:

5   a)   a digitalized voice signal is determined from the spoken language;

b)   a signal analysis ensues on the digitalized voice signal, feature vectors for describing the digitalized voice signal proceeding therefrom;

c)   a global search ensues for imaging the feature vectors onto a

10   language present in modelled form, whereby phoneme *each* ~~[sic]~~ of the language can be described by a modified hidden Markov model and each status of the hidden Markov model can be described by a probability density function;

d)   the ~~becomes [sic]~~ probability density function *is* adapted by

15   modification of the vocabulary in that the probability density function is split into a first probability density function and into a second probability density function; *and*

e)   the global search offers a recognized word sequence

· This arrangement is especially suited for the implementation of the

20   inventive method or of one of its developments explained above.

This [illegible] ~~Exemplary embodiments of the invention are presented and explained~~ ~~below with reference to the drawing.~~

~~Shown are:~~

~~Figure 1   an arrangement or, respectively, a method for recognizing spoken~~

25   ~~language.~~

Figure 1 shows an arrangement or, respectively, a method for the recognition of spoken language. The introduction to the specification is referenced for explaining the terms employed below.

In a signal analysis 102, a digitalized voice signal 101 is subjected to a

30   Fourier transformation 103 with following feature extraction 104. The feature vectors 105 are communicated to a system for global searching 106. The global search 106

considers both an acoustic model 107 as well as a linguistic model 108 for determining the recognized word sequence 109. Accordingly, the digitalized voice signal 101 becomes the recognized word sequence 109.

The phoneme inventory is simulated in the acoustic model 107 on the basis of hidden Markov models.

A probability density function of a status of the hidden Markov model is approximated by a summing-up of individual Gaussian modes. A mode is, in particular, a Gaussian bell. A mixing of individual Gaussian bells and, thus, a modelling of the emission probability density function arises by summing up a plurality of modes. A decision is made on the basis of a statistical criterion as to whether the vocabulary of the speech recognition unit to be recognized can be modelled better by adding further modes. In the present invention, this is particularly achieved by incremental splitting of already existing modes when the statistical criterion is met.

The entropy is defined by

$$H_p = - \int_{-\infty}^{\infty} p(\bar{x}) \log_2 p(\bar{x}) \, d\bar{x} \qquad (1)$$

given the assumption that $p(\bar{x})$ is a Gaussian distribution with a diagonal covariance matrix, i.e.

$$p(\bar{x}) = \mathcal{N}(\bar{\mu}, \sigma_n) = \frac{1}{\sqrt{(2\pi)^N}} \frac{1}{\prod_n \sigma_n} \cdot \exp\left(-\frac{1}{2} \Sigma_n \frac{(x_n - \mu_n)^2}{\sigma_n^2} \right) \qquad (2)$$

one obtains

$$H_p = \sum_{n=1}^{N} \log_2 \sqrt{2\pi e} \, \sigma_n \qquad (3),$$

whereby

μ references the anticipated value,

σ_n references the scatter for each component n, and

N   references the dimension of the feature space.

The true distribution $p(\bar{x})$ is not known. It is, in particular, assumed to be a Gaussian distribution. In the acoustic model, the probability $p(\bar{x})$ is approximated with

$$\hat{p}(\bar{x}) = \mathcal{N}(\bar{\bar{\mu}}, \sigma_n),$$

on the basis of random samples, whereby

$$\bar{\bar{\mu}} = \frac{1}{L} \sum_{l=1}^{L} \bar{x}_l$$

represents an average over L observations. The corresponding entropy as function of $\hat{\mu}$ is established by

$$H_{\hat{p}}(\hat{\mu}) = - \int_{-\infty}^{\infty} p(\bar{x}) \log_2 \hat{p}(\bar{x}) \, d\bar{x} \qquad (4),$$

which ultimately leads to

$$H_{\hat{p}}(\hat{\mu}) = H_p + \sum_{n=1}^{N} \frac{\left(\mu_n - \hat{\mu}_n\right)^2}{\sigma_n^2} \log_2 \sqrt{e} \qquad (5)$$

The anticipated value $E\left\{\left(\mu_n - \hat{\mu}_n\right)^2\right\}$ amounts to $\frac{1}{L} \sigma_n^2$, so that the anticipated value of $H_{\hat{p}}(\hat{\mu})$ is given as

$$H_{\hat{p}} = E\left\{H_{\hat{p}}(\hat{\mu})\right\} = H_p + \frac{N}{L} \log_2 \sqrt{e} \qquad (6).$$

Equation (3) thus derives for the entropy of a mode that is defined with a Gaussian distribution with a diagonal covariance matrix. The process is now approximated with an estimating. The entropy of the approximated process derives as

$$\hat{H} = H + \frac{N}{L} \log_2 \sqrt{e} \qquad (7).$$

The estimate is all the better the higher the number L of random samples is, and the estimated entropy $\hat{H}$ becomes all the closer to the true entropy H.

Let

$$p(\bar{x}) = \mathcal{N}\left(\bar{\hat{\mu}}, \sigma_n\right) \tag{8}$$

be the mode to be divided. It is also assumed that the two Gaussian distributions that arise as a result of the division process have identical standard deviations σ^s and are identically weighted. This yields

$$\hat{p}^S(\bar{x}) = \frac{1}{2} \mathcal{N}\left(\bar{\hat{\mu}}_1^S, \sigma^S\right) + \frac{1}{2} \mathcal{N}\left(\bar{\hat{\mu}}_2^S, \sigma^S\right) \tag{9}.$$

Given the assumption that $\mu_1 \approx \hat{\mu}_1$, $\mu_2 \approx \hat{\mu}_2$ and that μ_1 is at a sufficiently great distance from μ_2, the entropy of the split probability density function respectively derives as

$$\hat{H}^S = 1 - \sum_{n=1}^{N} \log_2 \sqrt{2\pi e} \ \sigma_n^S + \frac{1}{2}\left(\log_2 \sqrt{e} \ \frac{N}{L_1} + \log_2 \sqrt{e} \ \frac{N}{L_2}\right) \tag{10}.$$

As division criterion, a reduction of the entropy as a result of the split event is required, i.e.

$$\hat{H} - \hat{H}^S > C \tag{11},$$

whereby C (with C > 0) is a constant that represents the desired drop of the entropy. When

$$\frac{L}{2} = L_1 = L_2 \tag{12}$$

is assumed, then deriving as a result thereof is

$$\sum_{n=1}^{N} \log_2 \frac{\sigma_n}{\sigma_n^S} > \log_2 \sqrt{e} \ \frac{N}{L} + 1 + C \tag{13}.$$

One possibility of determining the mid-points of the two new modes is disclosed below. A preferred default is comprised in meeting the criterion for the

splitting. In the recited example, the value of $\hat{\mu}$ is allocated to $\hat{\mu}_1^S$. $\hat{\mu}_2^S$ receives a maximum likelihood estimate of those observations that are imaged onto $\hat{\mu}$ in the Viterbi path. These stipulations merely reveal one possibility without any intent of a limitation of the disclosed method to this possibility.

The following steps of the exemplary application shows the embedding into an arrangement for speech recognition or, respectively, a method for speech recognition.

Step 1:   Initialization: $\bar{\mu}_1^S = \bar{\mu}$, $\bar{\mu}_2^S = \bar{\mu}$.

Step 2:   Recognizing the expression, analyzing the Viterbi path;

Step 3:   For every status and for every mode of the Viterbi path:

     Step 3.1:   define σ_n;

     Step 3.2:   define L_2 on the basis of those observations that lie closer to $\bar{\mu}_2^S$ than to $\bar{\mu}_1^S$ and set $L = L_2$. If $\bar{\mu}_2^S$ and $\bar{\mu}_1^S$ are identical, then assign the second half to the feature vectors $\bar{\mu}_2^S$ and the first half to the feature vectors $\bar{\mu}_1^S$.

     Step 3.3:   correspondingly define σ_n^S on the basis of the L_2 expressions;

     Step 3.4:   Re-determine $\bar{\mu}_2^S$ on the basis of the average of those observations that lie closer to $\bar{\mu}_2^S$ than to $\bar{\mu}_1^S$;

     Step 3.5:   interpret division criterion according to Equation (13);

     Step 3.6:   if division criterion according to Equation (13) is positive, generate two new modes with the centers $\bar{\mu}_1^S$ and $\bar{\mu}_2^S$.

Step 4:   Go to step 2.

The following publications were cited in the framework of this document:

[1]   N. Haberland et al., "Sprachunterricht -- Wie funktionert die computerbasierte Spracherkennung?", c't -- Magazin für Computertechnik -- 5/1998, Heinz Heise Verlag, Hannover, 1998, pages 120-125.

5   [2]   C. H. Lee et al., "Speaker Adaptation Based on MAP Estimation of HMM Parameters", Proc. IEEE Intern. Conference on Acoustics, Speech and Signal Processing, pages II-588 through II-561.